Controlling a confound in predictive models with a test set minimizing its effect

Darya Chyzhyk[1, 2, *], Gaël Varoquaux[1], Bertrand Thirion[1], and Michael Milham[2, 3]

[1]Parietal - Inria / CEA. Paris-Saclay University, France
[2]Center for the Developing Brain, Child Mind Institute, New York, New York 10022, USA
[3]Center for Biomedical Imaging and Neuromodulation,
Nathan S. Kline Institute for Psychiatric Research, Orangeburg, New York 10962, USA
[*]Corresponding author: darya.chyzhyk@inria.fr

***Abstract*—Predictive models applied on brain images can extract imaging biomarkers of pathologies or psychological traits. Yet, a successful prediction may be driven by a confounding effect that is correlated with the effect of interest. For instance fluid intelligence is strongly impacted by age; age is well predicted from brain images; hence successful prediction of fluid intelligence from brain images might have captured nothing more than a biomarker of aging. Here we introduce a non-parametric approach to control for a confounding effect in a predictive model. It is based on crafting a test set on which the effect of interest is independent from the confounding effect. We name this strategy "*anti mutual-information subsampling*". We demonstrate the approach with a large sample resting-state fMRI and psychometric data of healthy aging subjects (n = 608). We show that using a linear model to remove the effect of age on the brain signals ("deconfounding") leads to pessimistic scores, as previously reported. Anti mutual-information subsampling does not require to remove from the brain signals the shared variance between aging and fluid intelligence, and hence does not display this pessimistic behavior. In addition, it is non-parametric and hence robust to violations of the linear hypothesis.**

***Index Terms*—confound, subsampling, phenotype, predictive models, biomarkers, statistical testing**

I. Introduction: confounded predictions

Large-scale imaging cohorts give increased statistical power, and make it possible to extract new imaging biomarkers that predict phenotypes such as neuropyschiatric pathologies [1] or individual traits [2]. Pattern-recognition methods can predict individual traits useful for clinicians to measure variations of the brain, healthy or pathological [3]. For instance, a reliable marker of brain aging can be extracted from both anatomical imaging and resting-state functional imaging [2]. However, predictive models that extract biomarkers can easily capture confounding effects. For instance, Power et al [4] showed that in-scanner head motion produces a significant confound for rest fMRI: motion creates systematic differences in brain signals, and in-scanner motion varies with subjects' age.

More generally, in cohorts of larger sizes and with more phenotypic informations systematic assessment of data quality is difficult; individuals are not recruited in homogeneous population with controlled criterion. It is then crucial to control that detected brain-behavior associations are not driven by unwanted effects. Indeed the presence of confounding variables that can mediate observed correlation or be a latent common cause to observations is a major impediment toward drawing conclusion from brain-behavior relationships [5].

Classical statistic analysis in brain imaging is based on the general linear model (GLM) [6], in which confounding effects are controlled by additional regressors to capture the corresponding variance. Such an approach shows limitations in predictive modeling settings. First, it is based on maximum-likelihood estimates of linear models, while predictive models are seldom maximum likelihood and sometimes non linear. Second, it is designed to control *in-sample* properties, while predictive models are designed for *out-of-sample* prediction. A two step approach based on applying a classical GLM to *deconfound* as a first step followed by a predictive model is often pessimistic, leading to below-chance prediction [2], [7].

In this paper, we develop a novel statistical framework to control for confounding effects in predictive models. This framework is based on *anti mutual information sampling*, a novel sampling approach to create a test set in which the effect of interest is independent from the confounding effect. This approach provides a non-parametric test of whether or not the imaging data drives a significant prediction, beyond the confounding effect. To demonstrate our approach, we consider prediction of fluid intelligence. Indeed, aging significantly impacts fluid intelligence [8], [9]. Given a fluid-intelligence predictor trained on a brain-imaging cohort with a wide age span, the CamCan dataset [10], the question is whether predictions are driven solely by age differences. Consistently with previous work, we find that a deconfounding approach yields pessimistic results. Our approach concludes that, on CamCan, brain signals predict fluid intelligence better than age, though not significantly.

II. Method: anti mutual-information subsampling

A. Formalizing the problem of prediction with a confound

We consider on n subjects: brain signals $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, an effect of interest $\boldsymbol{y} \in \mathbb{R}^n$ –the biomarker target– and a confounding effect $\boldsymbol{z} \in \mathbb{R}^n$. An imaging biomarker predicts $\boldsymbol{y}$ from $\boldsymbol{X}$ independently of $\boldsymbol{z}$. If $\boldsymbol{y}$ and $\boldsymbol{z}$ are not independent, we have to account for this effect. Indeed, a prediction of a target $\boldsymbol{y}$ mediated by a phenotypic information $\boldsymbol{z}$ may

978-1-5386-6859-7/18/$31.00

be misleading or useless. It can be misleading as it can be interpreted as a link between brain structures and $\boldsymbol{y}$ –e.g. fluid intelligence– while it is really mediated by $\boldsymbol{z}$ –e.g. age. It can be useless because brain imaging is likely much more costly to acquire than the phenotypic variable $\boldsymbol{z}$, hence may be preferable only if it bring mores diagnostic information. The problem that we focus on here it to *test* whether we can predict $\boldsymbol{y}$ from $\boldsymbol{X}$ rather than $\boldsymbol{z}$. Testing predictive model is done within a cross-validation loop, separating train and test sets [11]. Little et al [12] discuss what cross-validation measures in the presence of a confounding variable.

B. Existing approaches for predictions with confounds

a) Deconfounding: The classical procedure in the context of the GLM is to orthogonalize variables that are correlated [6]. In a *deconfounding* approach, a linear model can be used prior to the predictive model to remove the effect of the confounds $\boldsymbol{z}$ in the brain signals $\boldsymbol{X}$. It must be adapted to out-of-sample testing. One solution is to apply deconfounding jointly on the train and the test set, but it breaks the statistical validity of cross-validation [11] by coupling the train and the test set, hence it gives biased results. Another solution is to apply deconfounding separately on the train and the test set, but this can be very pessimistic [7]. Indeed the shared variance removed is then an *in-sample* property, and grows larger in an uncontrolled way for small test sets. The best approach is to consider the deconfounding model as a predictive encoding model, predicting a fraction of the signal $\boldsymbol{X}$ from $\boldsymbol{z}$, and removing it from $\boldsymbol{X}$. This model can be learned on the train data and applied to the test data.

Another drawback of deconfounding is that it is strongly parametric: it only takes into account second-order statistics (covariance or correlations) and may retain latent, more complex dependencies. A possible solution is to use a polynomial expansion of the confounds $\boldsymbol{z}$ in the deconfounding model.

b) Comparing predictive power: A simple safeguard to gage the impact of $\boldsymbol{z}$ on the prediction of $\boldsymbol{y}$ is to use predictive models predicting $\boldsymbol{y}$ from $\boldsymbol{z}$ and compare the predictive accuracy to that obtained with the biomarkers based on brain signals. Such argument is used by Abraham et al [1] to control for the effect of movement on autism diagnostic.

C. Creating a data subset to minimize mutual information

Our strategy is the following: we use as a test set a subset $\mathcal{S}$ of the data such that $\boldsymbol{y}_\mathcal{S}$ and $\boldsymbol{z}_\mathcal{S}$ are close to independent. The remainder of the data is used as a training set, to learn to predict $\boldsymbol{y}$ from $\boldsymbol{X}$. If the prediction generalizes to the test set $\mathcal{S}$, the learned relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$ is not entirely mediated by $\boldsymbol{z}$. In particular, the prediction accuracy then measures the gain in prediction brought by $\boldsymbol{X}$.

Technically, the challenge is to generate many test sets on which $\boldsymbol{y}_\mathcal{S}$ and $\boldsymbol{z}_\mathcal{S}$ are independent. Our approach is based on iterative sampling to match a desired distribution: our goal is independence, i.e. $p(\boldsymbol{y}, \boldsymbol{z}) = p(\boldsymbol{y})\, p(\boldsymbol{z})$, where $p((\boldsymbol{y}, \boldsymbol{z}))$ is the joint probability function of $\boldsymbol{y}$ and $\boldsymbol{z}$, and $p(\boldsymbol{y})$ and $p(\boldsymbol{z})$ are the marginal probability distribution. A related quantity is mutual information, which characterizes level of dependency between the two variables: $\mathbb{E}\left[\log\left(\frac{p((\boldsymbol{y},\boldsymbol{z}))}{p(\boldsymbol{y})p(\boldsymbol{z})}\right)\right]$. In practice we estimate the probability density functions with a kernel-density estimator (KDE) using Gaussian kernels. We iteratively create the test $\mathcal{S}$ set by removing subjects; at each iteration, we consider the problem as an distribution matching problem, matching $p(\boldsymbol{y}_\mathcal{S}, \boldsymbol{z}_\mathcal{S})$ and $p(y\mathcal{S})\, p(z\mathcal{S})$. For this, we use importance sampling: we draw randomly 4 subjects to discard with a probability $\frac{p(\boldsymbol{y}_\mathcal{S},\boldsymbol{z}_\mathcal{S})}{p(\boldsymbol{y}\mathcal{S})\, p(\boldsymbol{z}\mathcal{S})}$ using inverse sampling method [13, sec 2.2]. Algorithm 1 gives the details.

Algorithm 1: Anti Mutual Information Subsampling

Input: Target $\boldsymbol{y} \in \mathbb{R}^n$, confound $\boldsymbol{z} \in \mathbb{R}^n$, size $m < n$

$\mathcal{S} \leftarrow \{1 \dots n\}$ `/* Initialize */`

while *card*$(\mathcal{S}) > m$ **do**

  $p_y \leftarrow \text{KDE}(\boldsymbol{y}_\mathcal{S})$ `/* Density estimation */`

  $p_z \leftarrow \text{KDE}(\boldsymbol{z}_\mathcal{S})$

  $p_{(y,z)} \leftarrow \text{KDE}((\boldsymbol{z}_\mathcal{S}, \boldsymbol{y}_\mathcal{S}))$

  $\boldsymbol{m}_i \leftarrow \frac{p_{(y,z)}((\boldsymbol{z}_i, \boldsymbol{y}_i))}{p_y(\boldsymbol{y}_i, \boldsymbol{z}_i)},\ \forall i \in \mathcal{S}$

  $\mathcal{S} \leftarrow \mathcal{S} - \{j\}$ Draw one index j to remove from S with probability $\boldsymbol{m}_j$ using inversion sampling.

end

Output: Set of test indices $\mathcal{S}$

III. Empirical study

A. A rest-fMRI dataset with confounded prediction

Various studies have linked whole-brain functional connectivity to individual cognitive abilities [14], [15]. In particular, Hearne et al [15] predicted fluid intelligence from rest fMRI on the Human Connectome Project data.

We consider the Cambridge Centre for Ageing and Neuroscience (CamCan) data [10]. This study investigates brain anatomy and function as well as subjects' cognition over a large age span. Phenotype data comprises demographic (age) and psychometric (including fluid intelligence, measured by a Cattell test). Fluid intelligence is known to change during the lifetime, with a progressive decrease in middle age [16]. Indeed, the data displays a strong relation between fluid intelligence and age (fig. 1). When extracting biomarkers of fluid intelligence, the danger is to simply predict age. Below, we study how to control the impact of age.

We consider prediction from rest-fMRI functional connectivity: each row of $\boldsymbol{X}$ is a vectorized form of the functional connectivity matrix for each subject; the target vector $\boldsymbol{y}$ is the fluid intelligence score; the confound $\boldsymbol{z}$ is the age in years.

B. Experimental paradigm: cross-validating predictive models

a) Prediction pipeline: We use functional-connectivity matrices as brain imaging signals to build our biomarkers, following [1]. These matrices were generated by extracting the time series from predefined atlases BASC [17] with 64 regions and then calculating connectivity matrices using tangent [18] connectivity measure. We use Nilearn [19] to extract the signal

and build the matrices. As a prediction model we choose the standard ridge regression with nested cross-validation.

b) Experiments: We assess the prediction accuracy using a cross-validation procedure. To generate the test set, we consider two subsampling methods: either random subsampling, or the anti mutual information subsampling procedure presented above. The training set contains the samples that are not dropped to the test set. We use 100 folds. For the anti mutual information subsampling, different seeds in the pseudo random number generator lead to different fold. We assess the null distribution with permutations (20 000 folds on permuted labels $\boldsymbol{y}$). With random subsampling, we also consider deconfounding age.

C. Results

Fig. 1 displays joint distribution of the target (Fluid intelligence) and the confound (age) in the train and test set for different subsampling methods. The anisotropy of the cloud of points clearly reveals the statistical dependency between fluid intelligence and age. The test set generated by Anti Mutual Information method does not display such anisotropy, demonstrating the absence of such dependency.

For quantitative evaluation of such dependency, we plot in Fig. 2 the mutual information and correlation between the target and the confound for the different subsampling strategies as we progressively remove more subjects. With Anti Mutual Information subsampling, both decrease steadily until they hit a plateau at 300 subjects, which corresponds to a subset with independent age and fluid intelligence. In contrast, random sampling (with and without deconfounding) monotonically maintains high values of mutual information and correlation.

Finally, we report the prediction accuracy measured with the various approaches in Fig. 3. We use mean absolute error, as it gives an error measure in the unit of $\boldsymbol{y}$, comparable across different test sets. As expected, test sets generated by random subsampling lead to small prediction error on fluid intelligence, much better than chance-level prediction. However, deconfounding age leads to predictions that are much worst than chance. Forming a test set with anti mutual information subsampling avoids to create data where prediction of fluid intelligence is beyond chance. Instead, this prediction error is slightly better than chance, thought not significantly.

We hypothesize that deconfounding age leads to worst-than-chance prediction of fluid intelligence because it removes shared variance between age and fluid intelligence in the brain signals. To test this hypothesis, we use a control experiment where target and confound are the same variable: age. Fig. 4 shows that the effect of deconfounding is similar to that obtained when predicting fluid intelligence. This supports our hypothesis.

IV. Discussion and conclusion

We consider the problem of building biomarkers in the presence of confounding effects that contribute to prediction. Standard cross-validation procedures that do not account for the confound can overestimate the quality of prediction from brain signals. For instance, an observed prediction of fluid intelligence from brain functional connectivity may simply be a consequence of indirectly capturing the subjects age.

Deconfounding approaches used in standard GLM-based analysis must be adapted to out-sample data, and can easily lead to pessimistic evaluations [7]. Indeed, when there is shared signal between the effect of interest and the confound in the brain data, deconfounding fully removes this signal and makes it impossible to build a prediction from it. This is unlike a GLM-based standard analysis in which the confounds and the effects of interest are modeled *simultaneously*, and the consequences of shared signal are less drastic.

Our approach uses *anti mutual information* sampling to craft a test set on which the effect of interest is independent from the confound. It enables a correct test of the predictive power from brain imaging without killing potentially useful shared signal. In addition, it is non parametric and does not rely of a linear confounding model. We demonstrate the use of the method on data where aging confounds inter-subject variations in fluid intelligence and hinders the extraction of brain-based biomarkers of fluid intelligence.

V. Acknowledgment

This work was funded by the Child Mind Institute

References

[1] A. Abraham, M. P. Milham, A. Di Martino, R. C. Craddock, D. Samaras, B. Thirion, and G. Varoquaux, "Deriving reproducible biomarkers from multi-site resting-state data: An autism-based example," *NeuroImage*, vol. 147, pp. 736–745, 2017.

[2] F. Liem, G. Varoquaux, J. Kynast *et al.*, "Predicting brain-age from multimodal imaging data captures cognitive impairment," *NeuroImage*, vol. 148, p. 179, 2017.

[3] C.-W. Woo, L. J. Chang, M. A. Lindquist, and T. D. Wager, "Building better biomarkers: brain models in translational neuroimaging," *Nature neuroscience*, vol. 20, no. 3, pp. 365–377, 2017.

[4] J. Power, K. Barnes, A. Snyder, B. Schlaggar, and S. Petersen, "Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion," *Neuroimage*, vol. 59, p. 2142, 2011.

[5] J. Pearl, *Causality: Models, Reasoning, and Inference*. New York, NY, USA: Cambridge University Press, 2000.

[6] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-B. Poline, C. Frith, and R. S. J. Frackowiak, "Statistical parametric maps in functional imaging: A general linear approach," *Hum Brain Mapp*, p. 189, 1995.

[7] L. Snoek, S. Miletic, and H. S. Scholte, "How to control for confounds in decoding analyses of neuroimaging data," *bioRxiv*, 2018.

[8] J. K. Hartshorne and L. T. Germine, "When does cognitive functioning peak? the asynchronous rise and fall of different cognitive abilities across the life span," *Psychological science*, vol. 26, p. 433, 2015.

[9] D. Samu, K. L. Campbell, and et al., "Preserved cognitive functions with age are determined by domain-dependent shifts in network responsivity." *Nature communications*, vol. 8, May 2017.

[10] J. R. Taylor and et al., "The cambridge centre for ageing and neuroscience (cam-CAN) data repository: Structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample," *NeuroImage*, vol. 144, p. 262, jan 2017.

[11] G. Varoquaux and et al., "Assessing and tuning brain decoders: cross-validation, caveats, and guidelines," *NeuroImage*, vol. 145, pp. 166–179, 2017.

[12] M. A. Little and et al., "Using and understanding cross-validation strategies. perspectives on saeb et al." *GigaScience*, vol. 6, p. 1, 2017.

[13] L. Devroye, *Non-Uniform Random Variate Generation*. Springer-Verlag, 1986.

[14] E. S. Finn, X. Shen, and et al., "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature Neuroscience*, vol. 18, no. 11, pp. 1664–1671, oct 2015.

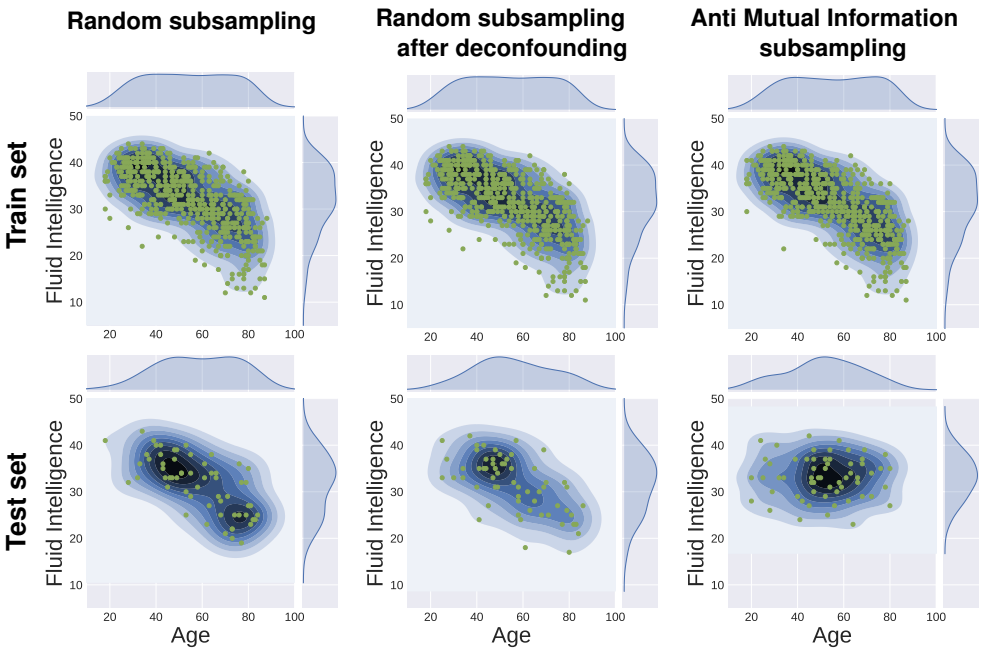

Fig. 1. **Bivariate distribution of Fluid intelligence and Age combined with distribution for both variables.** The top row shows the samples that were included in the training set. The testing set acquired by different sampled methods is presented on the bottom row. For the Random Sampling (first and second columns) we observe considerable similarity and negative covariance that indicate on strong relationship between predicted variable (fluid intelligence) and confound (age). Anti Mutual Information Subsampling method (third row) display zero covariance and interdependency between variables.

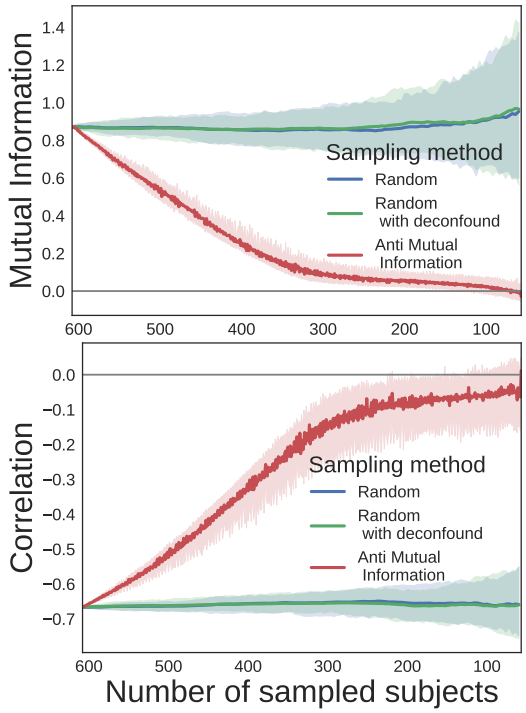

Fig. 2. **Evolution of mutual information and correlation with number of subjects for different subsampling methods.** We observe the decreasing of mutual information and tend to the zero of correlation from the initial set (608 samples) to final sampled set (around 60 samples).

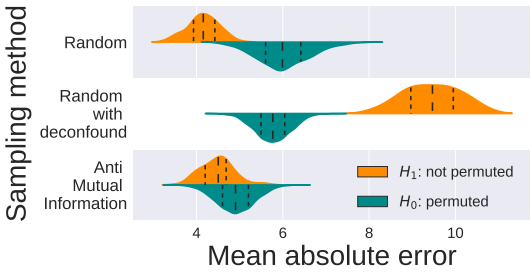

Fig. 3. **Comparison of the performances of subsampling methods on CamCan data set, predicting Fluid Intelligence.** The results of the prediction of fluid intelligence of sampling methods show, that random sampling capture the age information, the classical deconfounding gives significant but negative results. Proposed non-parametric Anti Mutual Information method is contrasting with unfavorable prediction. To demonstrate the significance of the presented discovers we show the results on the permutations (dark blue half of the violin plot).

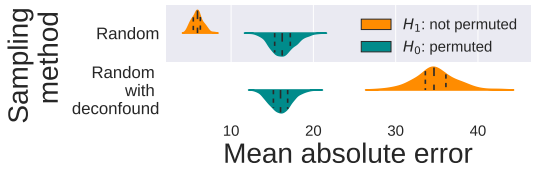

Fig. 4. **Predicting age, and deconfounding the effect of age.** As expected, the prediction of age is well below chance after removing the effect of age on the brain signals.

[15] L. J. Hearne, J. B. Mattingley, and L. Cocchi, "Functional brain networks related to individual differences in human intelligence at rest," in *Scientific reports*, 2016.

[16] J. L. Horn and R. B. Cattell, "Age differences in fluid and crystallized intelligence," *Acta Psychologica*, vol. 26, pp. 107 – 129, 1967.

[17] P. Bellec, P. Rosa-Neto, O. Lyttelton, H. Benali, and A. Evans, "Multi-level bootstrap analysis of stable clusters in resting-state fMRI," *NeuroImage*, vol. 51, p. 1126, 2010.

[18] G. Varoquaux, F. Baronnet, A. Kleinschmidt, P. Fillard, and B. Thirion, "Detection of brain functional-connectivity difference in post-stroke patients using group-level covariance modeling," in *MICCAI*, 2010.

[19] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in neuroinformatics*, vol. 8, 2014.